SI 11. Data processing workflow applied for untargeted UHPLC-MS studies

STEP 1: ACQUIRE (UP)LC-MS DATA IN PROFILE MODE

OUTPUT - SINGLE .RAW FILES FOR ALL SAMPLES



STEP 2: CONVERSION OF .RAW PROFILE MODE FILES TO mzML IN PROTEOWIZARD-MSCONVERT SOFTWARE OUTPUT – mzML FILES FOR ALL SAMPLES (CENTROID MODE)



STEP 3: PROCESSING OF mzML FILES IN A SINGLE BATCH APPLYING XCMS WITH 'OLD' MATCHED FILTER ALGORITHM OR 'NEW' CENTWAVE ALGORITHM AND CAMERA SOFTWARE.

OUTPUT - SINGLE .CSV FILE WITH INTEGRATED DATA



STEP 4: METABOLITE ANNOTATION APPLYING PUTMEDID LCMS

OUTPUT - LIST OF SINGLE OR MULTIPLE METABOLITES ASSIGNED TO A SINGLE FEATURE



STEP 5: BLANK SAMPLE ASSESSMENT. REMOVE OR FLAG FEATURES WHERE THE PEAK AREA (BLANK) > 5% OF PEAK AREA (MEDIAN OF ALL BIOLOGICAL SAMPLES)

OUTPUT – BLANK FILTERED DATASET CONTAINING ALL METABOLITE ANNOTATIONS



STEP 6: ASSESSMENT OF QC SAMPLES AND SIGNAL CORRECTION IF REQUIRED.

(SIGNAL CORRECTION IF BATCH SIZE >30 INJECTIONS).

- 1. REMOVE QC INJECTIONS 1-8 FROM DATASET
- 2. NORMALISATION OF DATA (SUM or PQN)
- 3. PCA OF ALL DATA, DO QC SAMPLES CLUSTER COMPARED TO BIOLOGICAL SAMPLES? IF SO CONTINUE.

 IF NOT STOP AND EVALUATE.
- 4. CALCULATE RSD ACROSS ALL BIOLOGICAL SAMPLES (BS) AND RSD ACROSS ALL QC SAMPLES FLAG
 ALL FEATURES WITH BS-RSD<2xQC-RSD
- 5. PERFORM SIGNAL CORRECTION APPLYING SPECIFIC ALGORITHM AND ORIGINAL NON-NORMALISED DATASET
- 6. RE-CALCULATE RSD ACROSS ALL BIOLOGICAL SAMPLES (BS) AND RSD ACROSS ALL QC SAMPLES FLAG ALL FEATURES WITH BS-RSD<2xQC-RSD. THIS NUMBER SHOULD BE GREATER THAN REPORTED IN STEP 4. IF SO CONTINUE. IF NOT STOP AND EVALUATE
- 7. FEATURE FILTERING, FLAG ALL FEATURES WITH RSD>20.0% FROM QC INJECTION 8, CHOOSE TO REMOVE ALL FEATURES WITH RSD>20% TO CONSTRUCT FILTERED DATASET OR FLAG FEATURES WITH RSD>20%
- 9. PCA OF FILTERED DATA, CLUSTERING OF QC SAMPLES? MOVE TO STEP 7 IF GOOD CLUSTERING.

 EVALUATE IF NOT GOOD CLUSTERING.OUTPUT BLANK AND QC FILTERED DATASET CONTAINING ALL

 METABOLITE ANNOTATIONS



STEP 7: FEATURE FILTER FOR EACH BIOLOGICAL CLASS.

REMOVE FEATURE IF 30% OR GREATER MISSING VALUES FOR ALL CLASSES
 REPLACEMENT FOR MULTIPLE CLASSES CONTAINS MORE THAN 70% MISSING VALUES AND ONE OR MORE OTHER CLASSES CONTAIN LESS THAN 30% MISSING VALUES THEN PERFORM SMALL VALUE REPLACEMENT FOR CLASSES WITH >70% MISSING VALUES (REPLACE ALL SAMPLES WITH SMALL VALUE OF LOWEST PEAK AREA IN SAMPLESET/2)

OUTPUT – BLANK, QC AND FEATURE FILTERED DATASET CONTAINING ALL METABOLITE ANNOTATIONS



STEP 8: SAMPLE CHECK. REMOVE QC SAMPLES, PERFORM ROBUST PCA, USE HOTELLIER'S T-SQUARED AND Q STATISTIC

OUTPUT – NORMALISED BLANK, QC AND FEATURE FILTERED DATASET CONTAINING ALL METABOLITE



STEP 9: UNIVARIATE ANALYSIS

PQN OR SUM NORMALISATION, RANDOM
FOREST MISSING VALUE IMPUTATION OR NO
MISSING VALUE IMPUTATION, NO
TRANSFORMATION OR SCALING

 TEST NORMALITY APPLYING SHAPIRO-WILK TEST

(if n>100) 3. PARAMETRIC TESTS – TTEST, ANOVA

4. NON-PARAMETRIC TESTS – MANN-WHITNEY,
KRUSKAL-WALLIS
5. FDR CORRECTED STATISTICAL ANALYSIS (APPLY

S. FDR CORRECTED STATISTICAL ANALYSIS (APPL BENJAMINO-HOCHBERG METHOD) 6. FOLD CHANGE WITH 95% CONFIDENCE INTERVALS



STEP 10: PCA ANALYSIS

RANDOM FOREST MISSING VALUE IMPUTATION,
 PQN OR SUM NORMALISATION, GLOG
 TRANSFORMATION, NO SCALING
 2. PCA

3. MULTIBLOCK PCA (CONSENSUS VS HCA)
OUTPUT – MULTIPLE DATAFILES CONTAINING
MULTIVARIATE ANALYSIS SCORES AND LOADINGS
PLOTS AND ALL METABOLITE ANNOTATIONS



STEP 11: PLS-DA ANALYSIS

1. PQN OR SUM NORMALISATION, BPCA+GLOG+RANGE METHODS OR KNN+GLOG+NO SCALING METHODS

2. PLS-DA

OUTPUT – MULTIPLE DATAFILES CONTAINING
MULTIVARIATE ANALYSIS SCORES AND LOADINGS
PLOTS AND ALL METABOLITE ANNOTATIONS